Creating a Web-Based Family History Questionnaire for Data Mining

Robert Hoyt MD FACP [1] Hui-Min Chung PhD [1] Brent Hutfless MS CISSP GSLC [1] Justice Mbizo Dr PH [1] Courtney Rice MS CGC [2]

[1] School of Allied Health and Life Sciences, University of West Florida, Pensacola, Florida
[2] Regional Perinatal Center, Sacred Heart Hospital, Pensacola, Florida

Abstract

As we move closer to ubiquitous electronic health records (EHRs) there will be a need to incorporate genetic, familial and clinical information as structured data that can be used for clinical decision support. While the Human Genome Project has produced new and exciting genomic data, the cost to sequence the human personal genome is high and significant controversies of how to interpret genomic data exist. Many experts feel that the family history is a surrogate marker for actual genetic studies and should be part of any paper-based or electronic health record (EHR). A digital family history is now part of meaningful use stage 2 requirements. It is likely that family history information will be used for clinical decision support in EHRs in the not too distant future. We designed an online family history questionnaire (FHQ) to collect computable data to determine whether resiliency (defined as no psychiatric illnesses post-trauma) in Vietnam-era repatriated prisoners of war (RPWs) is associated with any component of the family history. This paper describes our approach to create a digital FHQ in order to answer a research question.

Introduction

The Human Genome Project (HGP) accomplished whole genome sequencing in 2003 and resulted in voluminous data and new genetic tests. In spite of groundbreaking progress, the cost to perform whole genome sequencing remains high and experts are still debating many findings. According to the HGP web site "the scientific community continues to debate the best way to deliver them (results) to the public and medical communities that are often unaware of their scientific and social implications. While some of these tests have greatly improved and even saved lives, scientists remain unsure of how to interpret many of them". [1] Therefore, we are years away from incorporating genomic information into paper or electronic health records.

In the interim, many experts believe that information derived from a detailed family history will serve as a surrogate for personal genomic information.[2] Information derived from family histories is critical for focusing on medical disorders with a genetic component. [3-4] For example, national guidelines for the screening and management of cancer, diabetes and cardiovascular disease often include family history information. Unfortunately, obtaining detailed family

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
12 FEB 2013		**04-01-2012 to 12-02-2013**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Creating a Web-Based Family History Questionnaire for Data Mining	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
Robert Hoyt; Hui-Min Chung; Brent Hutfless; Justice Mbizo; Courtney Rice	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Robert E. Mitchell Center for Prisoner of War Studies,220 Hovey Road,Pensacola,FL,32508	**NMOTC-REMC-002**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
Navy Medicine Operational Training Center, 220 Hovey Road, Pensacola, FL, 32508	**NMOTC**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

Includes Sample Survey

14. ABSTRACT

As we move closer to ubiquitous electronic health records (EHRs) there will be a need to incorporate genetic, familial and clinical information as structured data that can be used for clinical decision support. While the Human Genome Project has produced new and exciting genomic data, the cost to sequence the human personal genome is high and significant controversies of how to interpret genomic data exist. Many experts feel that the family history is a surrogate marker for actual genetic studies and should be part of any paper-based or electronic health record (EHR). A digital family history is now part of meaningful use stage 2 requirements. It is likely that family history information will be used for clinical decision support in EHRs in the not too distant future. We designed an online family history questionnaire (FHQ) to collect computable data to determine whether resiliency (defined as no psychiatric illnesses post-trauma) in Vietnam-era repatriated prisoners of war (RPWs) is associated with any component of the family history. This paper describes our approach to create a digital FHQ in order to answer a research question.

15. SUBJECT TERMS

Electronic health records, information retrieval, questionnaires, genetics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**9**	

histories is time consuming, requires some expertise, [5] is not associated with clinician or patient reimbursement and is dependent on the quality of information provided by the patient. As a result, many paper-based records, personal health records (PHRs) and electronic health records (EHRs) fail to record vital family history that can be used to screen at-risk individuals and their families. Furthermore, rarely is the information captured as structured data, amenable to computation.

According to a 2009 systematic review by Reid et al. there have been about 14 family history questionnaires (FHQs) published in the literature but only 4 covered multiple medical conditions and none were web-based. Four studies were validated against a formal pedigree interview, considered the gold standard of genetic study. Nevertheless, the authors concluded that "there are no simple, short generic FHQs suitable for use in primary care practices". [6] New approaches to capture computable family history information must be taken.

Capturing family history information is important as studies have shown that most people are at moderate to strong risk of a medical condition with a genetic component.[7] However, this information is lacking in most current medical records. In spite of the increased adoption of electronic health records and personal health records, most commercial applications thus far do not include the ability to capture computable family history information. Data standards have been developed to help represent this important information, specifically, HL7 version 3 Clinical Statement Model and Clinical Genomics Family History Model, but significant limitations exist. [8]

The HITECH Act created reimbursement by Medicare and Medicaid for the meaningful use of certified EHRs in 2009. Stage 2 meaningful use, that is scheduled to go into effect in 2014, mandates that family history be recorded as structured data so it is computable.[9]

The federal government is aware of this void in family history information and has devised several projects to facilitate data collection. *My Family Health Portrait* is a free web-based open-source application patients and their families can use to create an electronic family history. This tool was developed by the Surgeon General and agencies within the US Department of Health and Human Services. Data standards used include HL7 Family History Model, LOINC and SNOMED CT. Patients can create a family history in about 15 minutes and then download the XML file to their computer or share it with their family or primary care physician. However, no interpretive features are available to evaluate the results for patients. It has been their hope that because this tool was written using data standards the results could be "consumed" by EHRs and PHRs. The software application is available as a download to individuals, healthcare organizations and developers.[10] As an example, data from My Family Health Portrait can be exported to RiskApps, a free application to identify and manage women at risk for hereditary breast and ovarian cancer.[11]

Another federal program, Family Healthware ™ is a web-based research tool developed by the Centers for Disease Prevention and Control (CDC) that can be used to assess a person's risk for six disease categories (coronary heart disease, stroke, diabetes, colorectal, breast and ovarian cancers). A personalized prevention message is generated from collected data. [12]

In addition to providing valuable clinical decision support to clinicians, data from FHQs can also be a valuable resource for research. Yarnell et al. was able to demonstrate that a family history of coronary heart disease and parental longevity are related, but independent of each other and common cardiovascular risk factors, in predicting future coronary events. [13] It is likely that

research will eventually "triangulate" information derived from clinical, genetic and family history data.

Our goal was to create a secure online digital family history questionnaire (FHQ) that would evaluate the hypotheses that resilient individuals (defined as no psychiatric illnesses post-trauma) have a family history associated with fewer psychiatric and/or medical illnesses, compared to non-resilient individuals.

Methodology

Participants: Our study population was comprised of 430 male repatriated prisoners of war (RPWs) from the Vietnam War era, as well as 118 comparison group subjects, matched for gender, age, education and combat roles in Vietnam. These individuals visit the Robert E. Mitchell Center for Prisoner of War Studies, located in Pensacola, Florida on a near annual basis. The program has been in existence since 1973 with some RPWs having 38 years of longitudinal physical and psychological data.[14] In spite of severe malnutrition, torture and solitary confinement, 57 percent of RPWs did not develop any evidence of psychiatric disease during the 37 years of follow up, while 43 percent developed psychiatric disorders, including posttraumatic stress disorder.[15] Under an Office of Naval Research funded grant we studied multiple predictive factors related to resiliency, defined broadly as the long term absence of psychiatric illness. This project is IRB approved and all patients signed a consent form.

Survey development: for content and face validity we convened an expert panel consisting of a PhD university-based Geneticist, a private genetic counselor, a neuropsychologist and an experienced Internal Medicine physician to decide on the appropriate survey design and selection of those common medical and psychiatric diseases with a genetic component. A literature review was also undertaken to determine the availability and relevance of existing FHQs. We also benchmarked our efforts with the recommendations made by the 2008 American Health Information Community's Family Health History Multi-Stakeholder Workgroup. [16]

We used a commercial survey instrument (SurveyMonkey) to create the web-based survey.[17] The survey has the following sections. The order of questions is displayed in figure 1 and as follows.

1. One: Demographic-type questions to include gender, adopted status, twin status and ethnicity, to be answered by all participants prior to proceeding.
2. Two: Personal health information divided into the following question categories. All categories have a free text optional answer option. The number of questions in each category is located in brackets. In this section only participants will include the age of diagnosis in a drop down menu.
    a. General condition questions (8)
    b. Heart condition questions (5)
    c. Cancer questions (14)
    d. Brain disease/neurodegenerative disease questions (6)
    e. Mental disorder related to learning disability questions (2)
    f. Mental disorder other than related to learning disability questions (8)
    g. Substance abuse questions (2)
3. Three: Mothers health

    a. Begins with living/deceased Y/N (drop down menu); current age or age of death (drop down menu), smoker status (drop down menu); served in military (drop down menu).
    b. The question categories 1-7 above are again asked (total of 50 questions) but there is no option to record age of diagnosis
    c. The survey questions are found in figure 2
4. Four: Father's health and is identical to the mother's health section.
5. Five: Sibling health and is identical to the mother's health section
6. Six: Children's health and is identical to the mother's health section

We chose to develop our own FHQ to answer specific research questions but have enough flexibility for other studies. Originally, we had hoped to modify the open source application *My Family Health Portrait*, but found several challenges. The current program does include psychological disorders but does not include nicotine or alcohol use. Furthermore, the categories for age of death ended at "60 years and older", which would not have been adequate to evaluate parental longevity. Lastly, as the program is written, each individual downloads their own family health portrait to their computer, making group data aggregation difficult. Our approach to create a customized FHQ was associated with numerous lessons learned, outlined in the following paragraphs.

The initial survey was pilot tested with 20 volunteers who provided valuable input resulting in modifications of the survey, such as including ex-smokers to the smoking history. During the design and testing of the familial history survey, a series of changes were made to streamline the user experience and improve data analysis and survey data reliability. The survey web service used, Survey Monkey, uses a linear progression model with question and distracter-level logic. The survey logic as enacted is designed to permit participants with few family members an expedient transition through the survey, answering only applicable questions. Conversely, the logic also allows the participant with a large family – up to 10 siblings and/or 10 children – the opportunity to respond for many family members while also limiting the overall time expenditure spent in the survey. Some of the challenges faced by the design team related to patient family history on an age-specific level. While it was deemed that knowledge of age at diagnosis was desirable, relying on participant memory for granular information relating to family members was not practical. Questions were added to improve functionality for the participants who were adopted, raised in single family homes or without knowledge of parents or additional siblings. This allows participants to skip over questions related to biological family members that are otherwise mandatory for survey completion. Other changes were made to reduce data duplication within the survey and in the resulting data set for analysis, which dropped overall variables from over 5000 to approximately 1300, a significant reduction in redundant data.

We elected to include information on first degree relatives only: proband/informant, children, siblings and parents in order to keep the FHQ concise and decrease the size of the resulting database. Also, the most common definition of a positive family history is the involvement of one or more first degree relatives.

Protected health information: several important privacy factors needed to be taken into consideration for designing and delivering the participant notification, including but not limited to: method of delivery, content of the message, point of distribution, and since the survey is

hosted online – anonymity of participants in regards to cookie permanence and Internet Protocol (IP) address logging at the point of data collection. A concerted effort was made to meet and exceed all of the expectations of research studies related to the HIPAA Privacy Rule, published by the Department of Health and Human Services. The method of delivery chosen was e-mail, a commonplace communication method and one adopted by a majority of the intended participants. The content of the message was an unassuming request for voluntary participation in an online survey. The message did not contain identifying information within the body of the message, and recipient email addresses were blind copied to reduce identification and the likelihood of unintended consequences in the improbable event that messages would be intercepted or sent in error. Included within the body of each message were two key elements for the survey; one, the Uniform Resource Locator (URL) used to gain access to the survey, and two, a unique identifier created for the sole purpose of this research. Lastly, the service used for hosting the web-based survey permitted anonymous data acquisition – allowing evaluators the ability to disable IP and email logging for survey participants.

The resulting data set is a de-identified collection of familial history results that is tied only to the unique identifier given to participants as part of the notification. The unique identifier serves as a key, as allowed by the Privacy Rule for research purposes, for re-identifying the participants during data analysis after survey completion. The identifier is created and stored within a protected document used for re-identification purposes. This identifier is a randomly chosen eight digit number that is sequenced for each participant at both the first and fourth place values, representing the ones and thousands digits respectively, which is done in order to eliminate confusion in the event of typographical errors during time of survey submittal. It is possible that this level of obfuscation would not be necessary with a limited data set and data use agreements, but the methodology employed prevents participant identification in the unlikely event that the survey host experiences a security breach and subsequent data exposure.

The intended survey targeted more than 450 subjects, thus a more automated method was required for formulating the e-mail and sending the messages from the Robert E. Mitchell Center for Prisoner of War Studies organizational email account. The technique chosen relied on a Microsoft Excel spreadsheet and an embedded macro. The macro was written in the Visual Basic language native to Microsoft Office products. When activated, the macro was designed to launch the local email client on a given host, populate a message with key elements such as email address and the unique identifier code, and release the message to the recipient. The content of this message in no way belied the purpose of the survey or the intention to collect familial medical history data. In spite of this, the chosen technique was seen as a preferable method to the one employed by the survey service, as having both survey results and originating data could constitute a method of re-identification by a third party.

Statistical Analysis: All active patrons of the Robert E. Mitchell Center were sent the FHQ via email starting in March 2011 and the collection period ended in June 2012. Several patrons elected to complete the FHQ while attending the Mitchell Center. A total of 448 FHQs were sent and 309 were completed for a completion rate of 70%. FHQ data was exported into SPSS version 18 for data analysis.[18] In our analysis we compared two groups: a resilient group and a non-resilient group. There were two categories of disease burden based on the FHQ: psychiatric (to include substance abuse) and medical. In order to develop a composite disease burden score we would need to factor in the size of the family. For example, if there are 8 first degree family

members and 2 have breast cancer; the score for that disease entity would be 2/8 or .25. Total score would be the sum of positive answers to the 50 questions, corrected for family size.

Total disease burden score = xA (depression) + xB (anxiety) + x C (PTSD)……..+ yA (diabetes) + yB (breast cancer)…… x = psychiatric, y = medical

Conclusion

Family history is an important part of any medical record and is a potentially valuable tool for disease prediction, prevention and research. We are moving towards genetic information being part of all medical record systems, but obstacles remain such as cost, immature data standards and the fact that electronic health records are not ready for input of this data. Family history information should also be available in all electronic health records, in a computable format, so clinical decision support tools can remind clinicians of important testing and risk assessment. Unfortunately, there is not a simple generic family history questionnaire (FHQ) available for common use in primary care, for use with or without an EHR.

We developed a web based FHQ as part of a research study to help evaluate resiliency in repatriated prisoners of war. Included are lessons learned to create a concise FHQ for research purposes. Further work is needed to determine and validate the optimal family history core questions, the best methods to collect this information, how to integrate computable family history information into EHRs, interoperable data standards and future clinical support tools.

Disclaimer

The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of Defense, nor the US Government or the University of West Florida. This research was sponsored by the Office of Naval Research grant #FHP-FY08-01.

References

1. Human Genome Project http://www.ornl.gov/sci/techresources/Human_Genome/medicine/medicine.shtml. Accessed January 5, 2011
2. Kmiecik T, Sanders D. Integration of Genetic and Familial Data into Electronic Medical Records and Healthcare Processes. February 2 2009. http://www.surgery.northwestern.edu/dos-contact/info-systems/Kmiecik%20Sanders%20Article.pdf Accessed January 10, 2011.
3. Rich EC, Burke W, Heaton CJ et al. Reconsidering the Family History in Primary Care. J Gen Intern Med 2004; 19:273-80
4. Guttmacher AE, Collins FS, Carmona RH. The family health history—more important than ever. N Engl J Med 2004:351:2333-6
5. Suther S, Goodson P. Barriers to the provision of genetic services by primary care physicians: a systematic review of the literature. Genet Med 2003;5:70-6
6. Reid GT, Walter FM, Brisbane JM, Emery JM. Family History Questionnaires Designed for Clinical Use: A Systematic Review. Pub Health Gen 2009;12:73-83
7. O'Neill SM, Rubinstein WS, Wang C et al. Familial risk for common diseases in primary care: the Family Healthware Impact Trial. Am J Prev Med 2009;Jun:36(6):506-14
8. Melton GB, Raman N, Chen ES et al. Evaluation of family history information within clinical documents and adequacy of HL7 clinical statement and clinical genomics family

history models for its representation: a case report. J Am Med Inform Assoc. 2010;17:337-40
9. Department of Health and Human Services. Federal Register September 4th 2012. Stage 2 Meaningful Use. http://www.gpo.gov/fdsys/pkg/FR-2012-09-04/pdf/2012-21050.pdf Accessed December 10 2012
10. My Family Health Portrait https://familyhistory.hhs.gov Accessed January 28, 2011.
11. Cancer Risk Assessment Software. RiskApps™. http://www.hughesriskapps.com/ Accessed February 4, 2011.
12. Family Healthware ™. Genomics Translation. Centers for Disease Control and Prevention. www.cdc.gov/genomics/famhistory/famhx.htm Accessed January 16, 2011.
13. Yarnell J, Yu S, Patterson C et al. Family history, longevity, and risk of coronary heart disease: the PRIME Study. Int Epidem Assoc 2003;32:71-77
14. Robert E Mitchell Center for Prisoner of War Studies. http://www.med.navy.mil/sites/navmedmpte/nomi/rpow/Pages/default.aspx Accessed January 16, 2011.
15. Segovia F, Moore J, Linville S, Hoyt R, Hain R. Optimism Predicts Resilience in Repatriated Prisoners of War: A 37-Year Longitudinal Study. J Trauma Stress 2012;25:1-7
16. Feero WG, Bigley MB, Brinner KM. New standards and enhanced utility for family history information in the electronic health record: an update from the American Health Information Community's Family Health History Multi-Stakeholder Workgroup. J Am Med Inform Assoc. 2008;15:723-8
17. SurveyMonkey www.surveymonkey.com Accessed January 4, 2011.
18. Statistical Package for the Social Sciences (SPSS) http://www-01.ibm.com/software/analytics/spss/ Accessed June 3 2012

Figure 1. Logic flow diagram

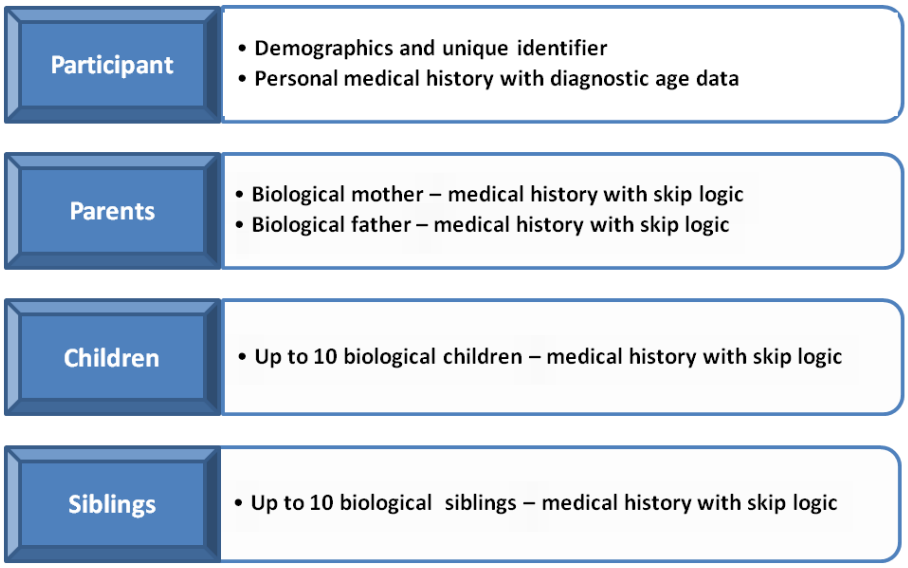
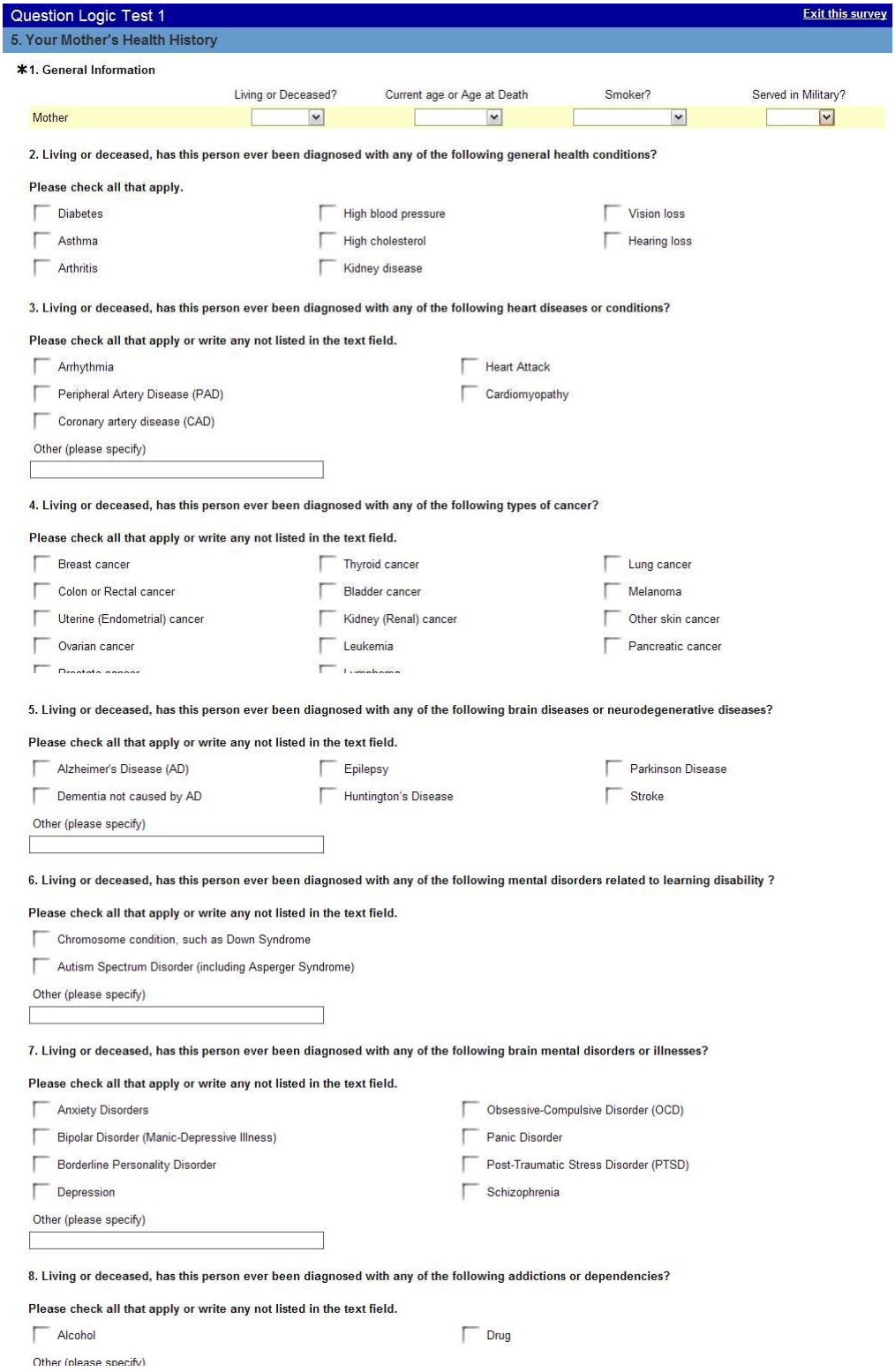


Figure 2. Sample Survey: Mothers Health History



Question Logic Test 1 — Exit this survey

5. Your Mother's Health History

*1. General Information

	Living or Deceased?	Current age or Age at Death	Smoker?	Served in Military?
Mother				

2. Living or deceased, has this person ever been diagnosed with any of the following general health conditions?

Please check all that apply.

- Diabetes
- High blood pressure
- Vision loss
- Asthma
- High cholesterol
- Hearing loss
- Arthritis
- Kidney disease

3. Living or deceased, has this person ever been diagnosed with any of the following heart diseases or conditions?

Please check all that apply or write any not listed in the text field.

- Arrhythmia
- Heart Attack
- Peripheral Artery Disease (PAD)
- Cardiomyopathy
- Coronary artery disease (CAD)

Other (please specify)

4. Living or deceased, has this person ever been diagnosed with any of the following types of cancer?

Please check all that apply or write any not listed in the text field.

- Breast cancer
- Thyroid cancer
- Lung cancer
- Colon or Rectal cancer
- Bladder cancer
- Melanoma
- Uterine (Endometrial) cancer
- Kidney (Renal) cancer
- Other skin cancer
- Ovarian cancer
- Leukemia
- Pancreatic cancer
- Prostate cancer
- Lymphoma

5. Living or deceased, has this person ever been diagnosed with any of the following brain diseases or neurodegenerative diseases?

Please check all that apply or write any not listed in the text field.

- Alzheimer's Disease (AD)
- Epilepsy
- Parkinson Disease
- Dementia not caused by AD
- Huntington's Disease
- Stroke

Other (please specify)

6. Living or deceased, has this person ever been diagnosed with any of the following mental disorders related to learning disability ?

Please check all that apply or write any not listed in the text field.

- Chromosome condition, such as Down Syndrome
- Autism Spectrum Disorder (including Asperger Syndrome)

Other (please specify)

7. Living or deceased, has this person ever been diagnosed with any of the following brain mental disorders or illnesses?

Please check all that apply or write any not listed in the text field.

- Anxiety Disorders
- Obsessive-Compulsive Disorder (OCD)
- Bipolar Disorder (Manic-Depressive Illness)
- Panic Disorder
- Borderline Personality Disorder
- Post-Traumatic Stress Disorder (PTSD)
- Depression
- Schizophrenia

Other (please specify)

8. Living or deceased, has this person ever been diagnosed with any of the following addictions or dependencies?

Please check all that apply or write any not listed in the text field.

- Alcohol
- Drug

Other (please specify)